

Generating and evaluating application-specific hardware extensions

NIKOLAOS KAVVADIAS

*Department of Computer Science and Technology, University of Peloponnese, Terma
Karaiskaki, Tripoli 22100, Greece
Email: nkavv@uop.gr*

Modern platform-based design involves the application-specific extension of embedded processors to fit customer requirements. To accomplish this task, the possibilities offered by recent custom/extensible processors for tuning their instruction set and microarchitecture to the applications of interest have to be exploited. A significant factor often determining the success of this process is the automation available in application analysis and custom instruction generation. In this paper we present YARDstick, a design automation tool for custom processor development flows that focuses on generating and evaluating application-specific hardware extensions. YARDstick is a building block for ASIP development, integrating application analysis, custom instruction generation and selection with user-defined compiler intermediate representations. In a YARDstick-enabled environment, practical issues in traditional ASIP design are confronted efficiently; the exploration infrastructure is liberated from compiler and simulator idiosyncrasies, since the ASIP designer is empowered with the freedom of specifying the target architectures of choice and adding new implementations of analyses and custom instruction generation/selection methods. To illustrate the capabilities of the YARDstick approach, we present interesting exploration scenarios: quantifying the effect of machine-dependent compiler optimizations and the selection of the target architecture in terms of operation set and memory model on custom instruction generation/selection under different input/output constraints.

Keywords: Embedded Systems; Hardware/Software Co-design; Performance Optimization; Electronic Design Automation; Application-Specific Processors

Received 00 December 2009; revised 00 Month 2010

1. INTRODUCTION

ASIPs (*Application Specific Instruction-set Processors*) play a central role in contemporary embedded systems-on-a-chip (SoCs) replacing hardwired solutions which offer no programmability for enabling reuse or encompassing late specification changes. ASIPs are tuned for cost-effective execution of targeted application sets. An ASIP design flow involves profiling, architecture exploration, generation/selection of instruction-set extensions (ISEs) and synthesis of the corresponding hardware while enabling the user taking certain decisions.

Custom processors either adhere to the configurable/extensible processor paradigm [1, 2] or can be ASIPs completely designed from scratch. Configurability lies in tuning architectural parameters (e.g. cache sizes) and enabling/disabling features [3] while exten-

sibility of a processor comes in modifying the instruction set architecture by adding forms of custom functionality. Designing a custom processor from scratch is a more aggressive approach requiring a significant investment of effort in developing all the necessary software development tools (compiler, binary utilities, debugger/simulator) and possibly a real-time OS, while in the configurable processor case, the RTOS is usually targeted to the base ISA and the software toolchain can be incrementally updated. There exist two basic themes for architecture extension: tight integration of custom functional units and storage [4] or loose coupling of hardware accelerators to the processor through a bus interface [5]. Recent works [6] advocate in favor of both approaches, proving that both techniques can be considered simultaneously by formalizing the problem as a form of two-level partitioning.

It is often in ASIP/custom processor design that

certain practical issues arising from seemingly invariant elements of the design flow are not addressed:

- a) Assumptions of the intermediate representation (IR) to which the application code is mapped, directly affect solution quality as in the case of ISE synthesis.
- b) The exploration infrastructure tied up to the conventions of software development tools.
- c) Adaptability to different compilers/simulators.
- d) Support for low-level entry for application migration within a processor family and reverse engineering.

In this paper, all these issues are successfully addressed by integrating custom instruction (CI) generation and selection techniques with a flexible IR infrastructure that can reflect certain designer decisions that is cumbersome to apply otherwise. Our approach is substantiated in the form of the YARDstick prototype tool [7]. For example using an IR with intrinsic support for bit-level operations may yield significantly different ISEs to the case of an unaugmented IR. Also, in YARDstick it is possible to directly measure the effect of certain machine-dependent compiler transformations, such as register allocation, to the quality and impact of the generated ISEs, an issue recognized but never quantified in other works [8, 9]. Further, YARDstick provides profiling facilities for determining static and dynamic application metrics such as data types, memory hierarchy statistics, and execution frequencies. Application entry can be either high-level (e.g. ANSI C) or low-level (assembly code for a target architecture or virtual machine). A number of recent custom functionality identification and selection techniques have been implemented while hardware estimators (speedup, area) and bindings to third-party tools for hardware synthesis from CDFGs are provided.

It is important to note that the interpretation of custom functionalities depends on the context; they can represent instruction-set extensions (ISEs) to a baseline ISA requiring to be accounted in the control path of the processor (decoding logic, extending the interrupt services), custom instructions of an ASIP enabled by a programmable controller or hardwired functions meant to be used as non-programmable hardware accelerators, loosely connected to the processor (i.e. accessible through the local bus).

2. RELATED WORK

Last years, a number of research efforts have regarded the automated application-specific extension of embedded processors [10, 11, 12, 13, 14, 15, 16]. A few open instruction generation frameworks exist [17]; an advantage of their work being delivering a format for storing, manipulating and exchanging instruction patterns. In order to use their pattern library

(Pattlib), the potential user should adapt his compiler for generating and manipulating patterns in the cumbersome GCC RTL (Register Transfer Language) [18] intermediate representation. Some issues with the Pattlib approach regard the significant efforts for adapting the GCC compiler to emit information in “pattlib” format, and that the IR for their selected backend (SPARC V8) is not architecture-neutral and cannot be easily altered.

Application-specific instructions have been generated for the Xtensa configurable processor [13] that may comprise of VLIW (Very Long Instruction Word), SIMD (Single-Instruction Multiple-Data) or fused (chained) RTL operations. However, as induced by the architecture template of Xtensa, control-transfer instructions (*cti*) are not considered to be included in the resulting complex instructions. A sophisticated framework for the design of tightly-coupled custom coprocessing datapaths and their integration to existing processors has been presented in [12]. While providing a complete solution to programmable acceleration, their work still has some drawbacks: the possibility of direct communication to fast local data memory is excluded and for this reason, beneficial addressing modes cannot be identified. In [19, 14] a multi-output instruction generation algorithm is presented which selects maximal-speedup convex subgraphs for each basic block data-dependence graph (DDG), with worst case exponential complexity, while [11] added path profiling to extend beyond basic block scope. An important conclusion was that useful instruction identification scope does not extend further than 2 or 3 consecutive basic blocks. Still, memory operations are not regarded in the formation of custom instructions, while pattern identification can only take place post register allocation.

3. YARDSTICK

The main role of YARDstick is to facilitate design space exploration (DSE) in heterogeneous flows for ASIP design where the development tools (compiler, binary utilities, simulator/ debugger) in many cases, lack DSE capabilities and/or have been designed with different interfaces in mind. Thus, it is often that significant development effort is required in adding features as afterthoughts and dealing with interoperability issues, especially at the *compiler* and *simulator* boundaries.

3.1. The YARDstick kernel

The current YARDstick infrastructure, as illustrated in Fig. 1, comprises of three kernel components (*libByoX*, *libPatCUTE*, *libmachine*), the target architecture specification tools (the BXIR frontend) and a set of backends for exporting control-flow graphs, basic blocks and custom instructions for visualization, simulation and RTL synthesis purposes. *libByoX* and *libPatCUTE*

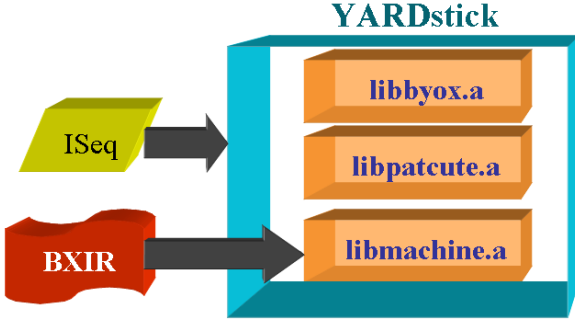


FIGURE 1. The YARDstick infrastructure.

are target-independent, and only *libmachine* has to be retargeted for different IR specifications.

3.1.1. *libByoX*

libByoX implements the core YARDstick API and provides frontends/manipulators for internal data structures. The ByoX (Bring Your Own Compiler and Simulator) library provides:

- The ISeqinfo parser for ISeq (historical name for “Instruction Sequence”) entries. ISeq is a flat CDFG (with/without SSA) format of application IR that is used for recording the data-dependence graphs for the application basic blocks.
- The CFGinfo parser for control-flow graph (CFG) files that attribute the corresponding ISeq files with typed control-flow edges.
- Simple file interface for the ISeq and CFG formats as well as for results of compiler analyses, e.g. control/data flow analyses evaluating register liveness and natural loops, that can be passed to ByoX as defined by their corresponding BNFs.
- An IR manipulation API for writing external analyses and optimizations.
- Parameterization for a template machine context without inherent restrictions to its ISA.

In ISeq, the following application information is recorded:

- The global symbol table.
- The procedure list, consisting of data dependence entries, the local symbol table and a statement list per procedure. It is possible to generate different facets of the local symbol table, e.g. single reference per direction (input or output) for each variable versus allowing multiple definition points for the same variable.

3.1.2. *libPatCUTE*

Further, a number of custom instruction generation/selection methods have been implemented as part of the PatCUTE (Pattern-based Custom UniT Exploration) library. CI generation involves the identification of MIMO (Multiple-Input Multiple-Output) or MISO

(Multiple-Input Single-Output) ISeq patterns under user-defined constraints. The CI generation methods available in *libPatCUTE* are:

- MAXMISO [10] for identifying maximal subgraphs with a single-output node using a linear complexity algorithm.
- MISO exploration under constraints for the maximum number of input/output operands, and for two types of operation node-related constraints [20].
- MIMO CI generation. In our case, we do not search for maximal MIMO patterns [16], however, we employ a fast heuristic by assuming similarly to [16] that the performance gain provided by a pattern P is higher than any pattern that is a subgraph of P . The user could disable the heuristic and apply an exponential complexity algorithm as well.

When CI generation is invoked, a CI list is constructed from the resulting ISeq patterns, which can be filtered via graph or graph-subgraph isomorphism tests [21] during the process of removing redundant cases. A subset of the library can be selected by using either a configurable greedy selector (supporting cycle-gain and cycle-gain per area priority metrics) or a 0-1 knapsack-based one. An important YARDstick characteristic is that CIs can be expressed in ISeq in the same way to either application CFGs or subregions thereof, thus existing data structures and analyses can be reused for further manipulation of the generated CIs. For example, pattern libraries can be imported to YARDstick.

3.1.3. *libmachine*

The *libmachine* library is the only core YARDstick component that needs retargeting for a user-defined target architecture. Target architectures are specified in the BXIR (ByoX IR) format which supports semantics for defining global-scope (data types, operation grouping) and operation-level information (operands, interpretation semantics for each IR operator, area/latency cost for corresponding hardware implementations and cycle timings).

3.1.4. *Backend engines*

Application CFGs, (basic blocks) BBs and patterns can be processed by a number of backends for exporting to:

- ANSI C subset code for incorporation to user tools (simulators, validators etc).
- GDL (VCG) [22] and dot (Graphviz) [23] files for visualization.
- An extended CDFG [24] format for scheduling and translation to synthesizable VHDL (applicable to BBs and CI patterns).
- GGX XML [25] files for algebraic graph transformation.

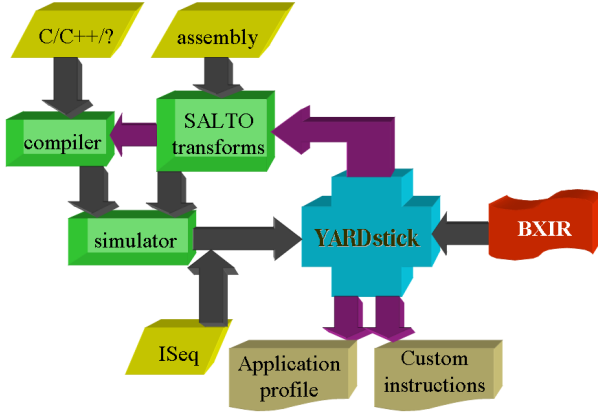


FIGURE 2. A high-level look to a YARDstick-based framework.

3.2. Structure of a YARDstick environment

The YARDstick kernel can be utilized as an infrastructure for application analysis and exploration of custom functionality extensions. Fig. 2 shows our YARDstick framework which reuses third-party compilation and simulation tools. The compiler frontend (*gcc* is such an example) accepts input in C/C++ or other high-level languages of interest. The application program is compiled to a low-level representation that can be represented by a form of “assembly” code after frontend processing, conversion to its internal IR, application of machine-independent optimizations and a set of compiler backend processes with only code selection being obligatory. The assembly-level code can then be macro-expanded, instrumented for profiling and converted to ISeq by an appropriate SALTO pass [26]. This flow assumes that a working SALTO backend library has been ported for the target architecture. Assembly code can be assembled and linked by the target machine binary utilities (*binutils* or equivalent tools) and the resulting ELF executables can be evaluated on an instruction- or cycle-accurate simulator. Alternatively, ISeq files can be generated as compiler IR dumps directly from the compiler for the target machine. This is the case for a modified version of Machine-SUIF [27] for which the basic block profile is automatically obtained by converting the IR to a C subset and executing the low-level C code on a native machine.

At the simulation boundary, YARDstick expects information on the dynamic profile of the application (basic block execution frequencies, program trace, cache memory access statistics) on a target machine. From within YARDstick, static and dynamic application metrics can be evaluated and visualized. An application analyzer (*iseqtool*) and CI generator (*igensel*) linked to *libByoX* and *libPatCUTE* are used to obtain the application profile and custom instructions, respectively.

```
void evaluate_bb_ci(ISeq bb)
{
    ...
    // Setup operand list
    UIOCList Lopnd = init_opnd_library();

    // Find unique i/o registers and constants
    find_input_opnds(bb, Lopnd, input_opnds);
    find_output_opnds(bb, Lopnd, output_opnds,
        instr_has_successor);
    find_cnst_opnds(bb, Lopnd, cnst_opnds);

    if (unique i/o instances for operands/constants)
        collapse_to_unique_opnds();

    clear_best_cut();

    // CI generation for the BB
    if (MIMO method)
        MIMO_identification(bb);
    else if (MaxMISO or constrained MISO method)
        MAXMISO_identification(bb);
}
```

FIGURE 3. Updating internal data structures for BB-level CI generation.

3.3. Usage of the YARDstick API

The YARDstick API provides methods for manipulation of ISeq entities and extraction of useful information to internal data structures such as local operand lists, operation-level analysis (e.g. finding zero-predecessor/-successor instruction nodes) and application of backend processing. Fig. 3 shows an example of API usage for updating the necessary data structures for basic-block-based CI generation.

In more detail, a basic block ISeq cluster is denoted by ‘bb’. First, ‘init_opnd_library’ initializes an empty operand list container, named *Lopnd* which is updated by calls to the ‘find_type_opnds’ functions, where *type* denotes operand type and can be one of {input,output,cnst}. When the unique register operands and constants option is enabled, operands are treated as in SSA form and have a single representation per input and output sublist by applying ‘collapse_to_unique_opnds’. After clearing temporary storage for the best cut to be identified in the specific iteration of the CI generation algorithm, either a MIMO or a MISO-based method can be selected for performing the actual process.

4. CASE STUDIES OF DESIGN SPACE EXPLORATION WITH YARDSTICK

For proof-of-concept, we have evaluated YARDstick under various scenarios that reflect realistic problems in evaluating and exploring the design space when developing new ASIPs or enhancing customizable architectures. For the case studies we have used three different target architectures:

TABLE 1. Different IR settings for CI generation.

IR	Operations
SUIFvmenh	SUIFvm plus: type conversion (sxt, zxt), partial predication (select), bit manipulation (bitinsert, bitextract, concat)
SUIFrmenh	SUIFvmenh with finite register set (12, 16, 32 or 64 registers); here 32 is used
iDLX	The DLX integer instruction set

TABLE 2. Summary of examined benchmarks.

Benchmark	Description
<i>crc32</i>	Cyclic redundancy check
<i>deraiden</i> [28]	Decoding raiden cipher
<i>enraidn</i> [28]	Encoding raiden cipher
<i>idea</i>	IDEA cryptographic kernel
<i>sha</i>	Secure Hash Algorithm producing an 160-bit message digest for a given input
<i>adpcmdec</i>	Adaptive Differential Pulse Code Modulation (ADPCM) decoder
<i>adpmenc</i>	Adaptive Differential Pulse Code Modulation (ADPCM) encoder
<i>fir</i>	FIR filter
<i>fsme</i>	Full-search motion estimation
<i>mc</i>	Motion compensation

- 1) The SUIFvm IR [27] augmented by a set of incremental extensions to it, called *SUIFvmenh*.
- 2) The *SUIFrmenh* architecture (SUIF ‘real machine’ enhanced) supported by an in-house backend written for Machine-SUIF, that introduces a finite register set of configurable size to *SUIFvmenh*. *SUIFrmenh* also resolves type casting (conversion) operations mapping them from the *cvt* SUIFvm instructions to the proper instructions accepted by the *SUIFrmenh* backend: zero- and sign-extend, truncation and *mov* explicitly denoting the source and destination data types.
- 3) The DLX integer subset (*iDLX*) for which the formatted assembly dumps are viewed as a kind of human-readable machine-level IR.

The target IR architectures are summarized in Table 1. In the experiments of the following subsections, all control transfer instructions (*beqz*, *bnez*, *j*, *jr*, *jal*, *jalr*) were forbidden from CI pattern formation for the *iDLX* IR, while for the SUIFvm-based IRs, branch operations were permitted. The two different forbidden instruction constraint sets were chosen in order to highlight distinct potential requirements and were not meant to be directly contrasted. The DLX-based IR would be a choice when the objective is to optimize pre-existing DLX legacy assembly (or binaries). Instead, *SUIFvmenh* implements a representative RISC-like IR not restricting the processor template, which is more suitable for developing ASIPs from scratch.

For the experiments we used applications from a set of embedded benchmarks consisting of 5 cryptographic (*crc32*, *deraiden*, *enraidn*, *idea*, *sha*) and 5 media-oriented applications (*adpcm_dec*, *adpcm_enc*, *fir*, *fsme*, *mc*) which are shown in Table 2.

4.1. Effect of compilation specifics: Case study of media processing kernels

It has been argued recently [29] that traditional compiler transformations and the trivial solution of applying CI identification at the end of the optimization phase pipeline do not necessarily yield the best performance when targeting a custom processor. On the contrary, source code and IR-level transformations have to be especially tuned for exposing beneficial application-specific hardware extensions.

In this subsection, the effect of the choice in compilation specifics is highlighted for popular case study applications: the ADPCM codec, an FIR filter, and typical implementations of motion estimation/compensation. We investigate specific effects that the compiler imposes when used for exploring the potential for custom instructions:

- a) The effect of register allocation on the quality of the generated CIs. For this purpose, we have targeted the *SUIFrmenh* backend. A 32-entry register file was assumed while the procedure calling convention for *SUIFrmenh* was the same to an in-house GCC-based DLX backend.
- b) The suitability of using a highly-optimizing (but aimed to general-purpose processors) compiler such as *gcc* targeted to DLX which is our case, against a well-known research compiler (*MachSUIF* targeted to SUIFvm) which has been extensively used for exploring the transformation space for new CIs.

For accounting only the true data dependencies amongst operations, it is necessary to remove all false dependencies. This can be achieved by a simple IR-level transformation pass (for example, such pass was implemented in the Machine-SUIF compiler for the *SUIFvmenh* target) which involves the use of the pseudocode of Fig. 4. The algorithm in Fig. 4 can be used for an in-order instruction schedule, i.e. no backward data dependence edges exist within a basic block. For a given set of dependence edges $\bigcup\{(i \rightarrow j)_k\}$ between instructions $mi(i), mi(j)$ of instruction IDs i, j respectively, the range $[i, j]$ is considered. The destination operands of machine instructions in range are iterated and compared to the operand (*opnd*) for which we want to remove all false dependencies with it as the data dependency. If *opnd* is written at least once, a false dependency is identified (marked as *TRUE*) and the corresponding data dependence edge is annulled.

Application speedups obtained prior and post register allocation (the latter indicated by a ‘ra’ suffix to the benchmark name) are shown in Fig. 5. In contrast to common belief [8, 9], the introduction of a finite register set and the mapping of the instruction selection temporaries to this set, does not always have a negative impact on the evaluated speedups. While it is clear that there is a measurable effect (an overhead of 22.15%) due to register allocation for a single output ($N_o =$

```

boolean is_false_dependency(BB* bb, InstrID mi_lpos,
    mi_hpos, LOpnd opnd)
{
    boolean false_dependency_f = FALSE;
    ...
    // Iterate through the [mi_lpos..mi_hpos] range
    foreach machine instruction (mi) in range do
        if the current mi is within the specified range
            get destination operand dstop of mi
            if dstop is ((a base register or address symbol)
                and writes memory)
                if dstop is equal to opnd
                    // a false dependency has been found
                    false_dependency_f |= TRUE;
            fi
        fi
    fi
    return false_dependency_f;
}

```

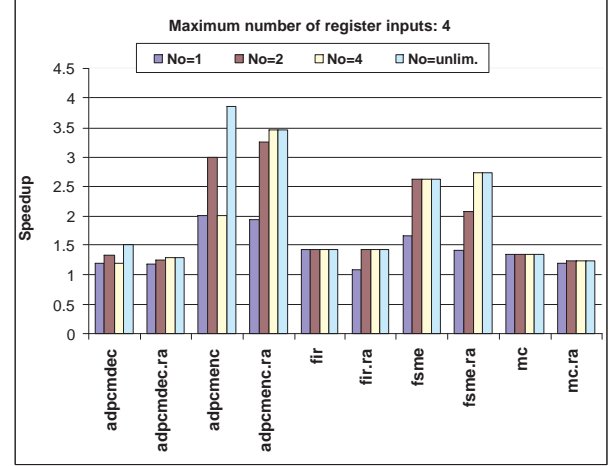
FIGURE 4. Removing false data dependence edges from basic blocks.

1), the extent of this overhead is reduced for larger number of register outputs. Thus, for $N_o = \{2, 4, \infty\}$ the corresponding overheads have been calculated as 17%, 2.5% and -21.3%, the latter meaning that the register allocated IR enables higher speedups compared to obtaining the IR prior register allocation for the constraint of unlimited number of register outputs. This important outcome infers that the overhead of spills and fills occurring due to register pressure, can be efficiently hidden when multi-output (MIMO) instructions are used for the estimations. In addition to that, CIs have the side-effect of eliminating the need for certain temporary variables within a CI pattern, given that they need not be alive outside the pattern.

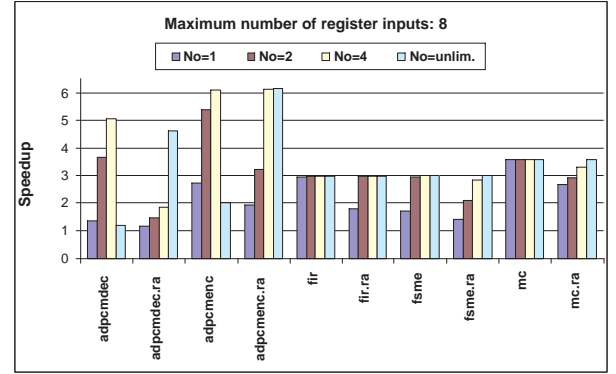
Fig. 6 shows the results on relative application speedups for different number of input/output register operands for the two selected compiler targets. An unlimited number of inputs was also set but the corresponding results were within 0.4% of the $N_i = 8$ case. The difference in the average speedup achieved for the given numbers of inputs for the same application is about 44% (ranging from 20% to 61%). This is partially due to the fact that stack argument allocation applied for *iDLX* only, adds memory access operations for saving and restoring function arguments that are not usually included in new CIs. Even when MIMO instructions are identified incorporating the callee saved sequence (a series of *sw* instructions), the obtained speedups are severely limited by the data memory bandwidth assumed in the estimations which is 1R/1W port for all target architectures.

4.2. Transformation to more suitable IRs

Although not explicitly stated in related works, the effect of IR selection significantly affects the quality of the CI generation results. In YARDstick, GGX



(a) $N_i = 4$

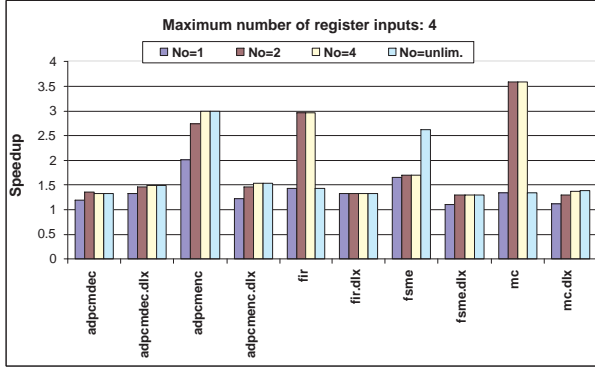
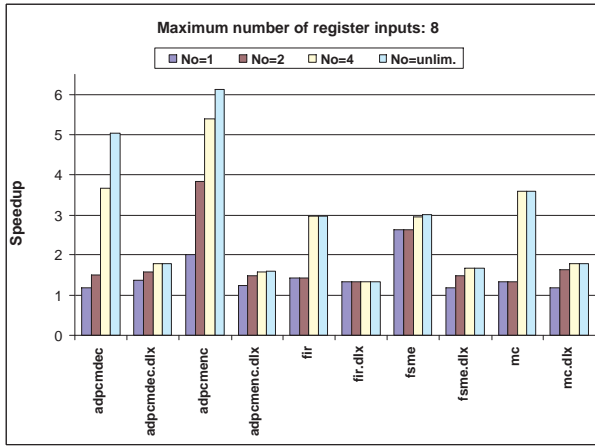


(b) $N_i = 8$

FIGURE 5. Effect of register allocation on application speedup for different number of input/output register operands.

XML graph representations of ISeq patterns can be automatically generated and then transformed by hand-written AGG rules to use different IR operators for implementing equivalent functionality.

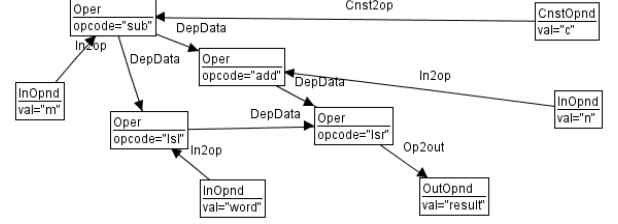
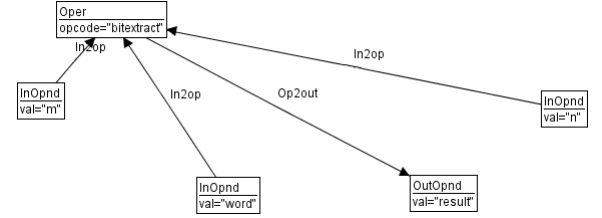
Most compilers (one exception is the commercial CoSy [30]) do not account for bit-level manipulations that are desirable in application domains such as network processing and genetic algorithms (GAs). To highlight this issue we have defined three custom IR operators, namely *bitinsert*, *bitextract* and *concat* with the semantics of Table 3. As motivational examples, we have used the well-known single- (*crcsp*) and double-point (*crcdp*) crossover operators, which are encountered in typical GAs such as the SGA [31]. It should be noted that the ANSI C implementations of the crossover operators were hand-tuned, with optimizations including the conversion of all function calls inside the *crcsp* and *crcdp* functions to macro-inclusions. Fig. 7 shows the result of applying a rule-based transformation in AGG [25] for replacing a *SUIFvmenh* IR segment (Fig. 7(a)) with a use of the *bitextract* IR operator as seen in the resulting graph (Fig. 7(b)). To highlight the importance of the

(a) $N_i = 4$ (b) $N_i = 8$ **FIGURE 6.** Application speedup for different number of input/output register operands on *SUIFvmenh* and *iDLX*.**TABLE 3.** Custom IR operators improving bit-level compiler support. r_d, r_s are register operands, $hpos, lpos$ denote a bit range, and n is the number of arguments for a variadic operator.

Operator	Semantics
<i>bitinsert</i>	$r_d[lpos..hpos] \leftarrow r_s$
<i>bitextract</i>	$r_d \leftarrow r_s[lpos..hpos]$
<i>concat</i>	$r_d \leftarrow r_{s(0)} \& r_{s(1)} \& \dots \& r_{s(n-1)}$

right choice of compiler IR, Fig. 8 visualizes the VCG representations of the custom instruction generated for the *crcsp* genetic operator, without (Fig. 8(a)) and with the use of the bit-level IR operators (Fig. 8(b)).

The performance gains for the generated hardware depend heavily on the target IR used for mapping the application code as can be clearly seen by the results of Table 4. In Table 4, the first three columns are self-explanatory. Column ‘Cycles...’ shows the cycles required for a sequential schedule of the corresponding GA operator assuming the usage of the generated CIs. The last two columns indicate the number of cycles and area of the CI. The area requirement is calculated relatively to the area (multiplier area unit or MAU) of a 32-bit single-cycle multiplier producing a 64-bit result. For computing schedules with unlimited resources,

(a) Visualization of an example host *SUIFvmenh* IR graph.

(b) The resulting graph after the application of a transformation rule for ‘bitextract’.

FIGURE 7. An example of IR graph rewriting via AGG transformation rules.**TABLE 4.** CI characteristics for hand-optimized ANSI C implementations of *crcsp* and *crcdp*.

GA operator	Bit-level operations	CI gen. constraints N_i/N_o	Cycles (seq. schedule)	CI cycles	CI area (MAU)
<i>crcsp</i>	No	4/1	76	—	—
<i>crcsp</i>	No	8/1	41	3	0.977
<i>crcsp</i>	No	8/2	5	3	1.867
<i>crcsp</i>	Yes	4/1	13	—	—
<i>crcsp</i>	Yes	8/1	6	1	0.142
<i>crcsp</i>	Yes	8/2	1	1	0.153
<i>crcdp</i>	No	4/1	111	—	—
<i>crcdp</i>	No	8/1	58	3	1.466
<i>crcdp</i>	No	8/2	5	3	2.800
<i>crcdp</i>	Yes	4/1	18	—	—
<i>crcdp</i>	Yes	8/1	8	1	0.147
<i>crcdp</i>	Yes	8/2	1	1	0.164

the generated ISeq files of the custom instructions were automatically converted with YARDstick to CDFGs compatible with an extended version of the CDFG toolset [24] and processed by an ASAP scheduler. If the bit-level operators are not used, the minimum number of cycles required for the *crcsp* operator are 76 for a sequential schedule and 12 for scheduling with unlimited resources, while for the *crcdp* these limits are 111 and 14, respectively. When the bit-level operators are used, the sequential schedules prior to the inclusion of custom instructions require 13 and 18 cycles for *crcsp* and *crcdp* respectively with an ASAP schedule of 5 cycles for both. In the latter case, a single-cycle MIMO custom instruction is identified for each genetic operator when a $N_i/N_o = \{8/2\}$ constraint is used with impressive area benefits as well.

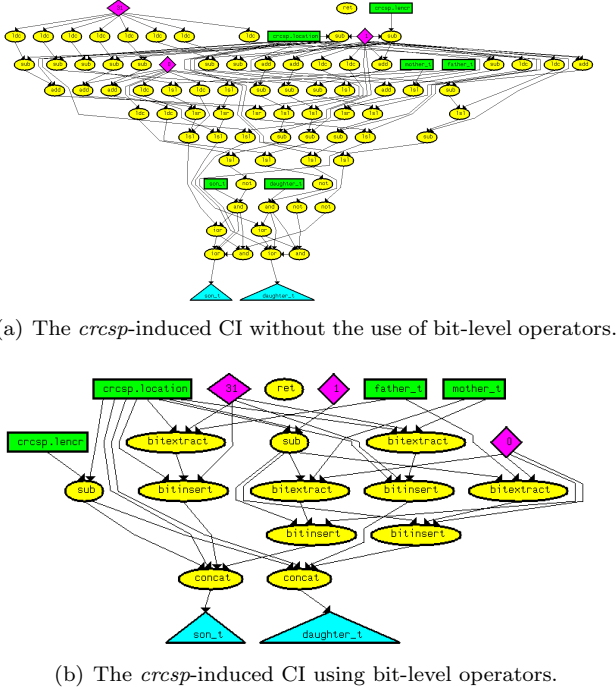


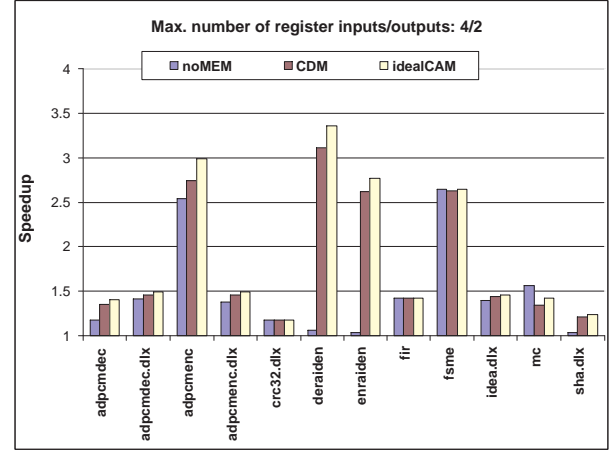
FIGURE 8. Visualization of the *crcsp* genetic operator CI for different compiler IRs.

4.3. Effect of data memory access model

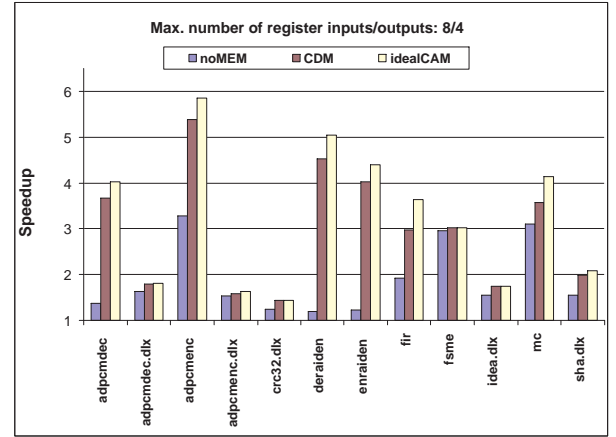
The extent and scope of using custom instructions is constrained by the data bandwidth to the data memory unit and local storage (register file) as defined by the number of input/output ports and the resolution of dependencies for load/store operations. In certain approaches [12, 32] which deal with predefined architectures such as the MIPS CorExtend and ARM OptimoDE systems, this limitation imposes a definitive factor. However, for exploration purposes when developing an ASIP from scratch, it is useful to consider different storage consistency models. Following the notation introduced in [15] for state consistency between application-specific functional units (AFUs) with local storage and data memory, it is possible to consider two such models in YARDstick:

- *Consistent data memory*, where the AFU directly accesses data in the on-chip data memory and there is no need for local AFU storage. We also make the conservative assumption that loads and stores ought always be serialized.
- *Ideal consistent AFU memory*, where each load/store to main memory is transformed to an access to local AFU memory. We assume that data memory status is updated by DMA accesses occurring in parallel to processor instructions.

To investigate the effect of memory model choice on application speedup due to CIs we first generated CIs without allowing memory inclusion (“noMEM”), then allowed local memory and performed estimations



(a) $N_i/N_o = 4/2$



(b) $N_i/N_o = 8/4$

FIGURE 9. Effect of data memory accesses to the speedup induced by CIs. Accesses to data memory are assumed to require a single clock cycle overhead.

for the consistent data memory model (“CDM”) and subsequently we assumed an ideal consistent AFU memory (“idealCAM”). The corresponding results are illustrated in Fig. 9 for indicative N_i/N_o combinations and for a single-issue processor.

As can be seen by the collected results, the inclusion of data memory access operations in CIs has a significant positive impact in the attained speedups: from 15.5% to 33.3% for the given input/output constraints. Especially for the *SUIFvmenh* target, the speedup improvements were up to 43.4%. Another important observation is that the consistent AFU memory model has a limited effect with improvements of up to 6.3% in average and 8.9% for the *SUIFvmenh* applications alone. However, for a larger cycle overhead to accessing data storage, the speedup improvements are more considerable. For another exploration example, we have estimated that 2- and 5-cycle load/store accesses to a data memory module through the local bus (an address cycle followed by either one word data access or consecutive byte data access cycles)

result in higher speedups. More specifically, the “CDM” case performs better to “noMEM” by 34.7% and 49.9%, respectively for the 2- and 5-cycle overheads. When comparing the two different models that allow memory accesses to be part of CIs, the corresponding values are 7% and 20.9% in favor of “idealCAM” for the given cycle overheads.

4.4. Greedy CI selection under priority metrics

For implementing a greedy CI selector, the key idea is to assign priorities to the CI patterns and the more proficient instances are chosen by starting with the highest prioritized one. We have used the following two priority functions:

$$\text{Cycle gain : Priority}(\sum_j C_{i,j}) = \sum_j \{P_{i,j} \times f_{i,j}\} \quad (1)$$

that forces for best performance regardless AFU area requirements and:

$$\text{Cycle gain/Area : Priority}(\sum_j C_{i,j}) = \sum_j \{(P_{i,j} \times f_{i,j})\} / A_i \quad (2)$$

where $C_{i,j}$ denotes the i -th candidate instruction with j different instances in the entire program, $f_{i,j}$ the basic block execution frequency metric associated with the specific instance, and A_i the area cost for the candidate. These priority functions force different objectives: equation 1 maximizes performance gain for each isomorphic candidate CI over the entire program when area is not an issue while equation 2 quantifies the available area budget as well.

A summary of the measurements for the application set is given in Table 5. Taking *sha.dlx* for example, although tens of candidate instructions are identified, only a few (7 for achieving 95% of the maximum speedup compared to 20 for achieving totality) contribute significantly to the execution time for either priority function. The number of required extensions for reaching the 95% speedup levels ranges from 2 (*fir.dlx*) to 40 (*idea.dlx*), while the area requirement is less than 3.4 multiples of the area of a 32-bit single-cycle multiplier for all applications with the exception of *idea.dlx* which demands up to 10.23 MAU.

Finally, Fig. 10 compares the pros and cons for the priority functions used in the custom instruction selection process for the *sha.dlx* application example. For the *sha* application, CI selection under the ‘Cycle gain’ priority function reaches the 95% of the maximum speedup for one instruction less and a slight area increase (0.04 MAU) compared to ‘Cycle gain/Area’.

5. USAGE ENVIRONMENT

YARDstick has been used along with the SUIF/Machine-SUIF [27], GCC [18], and COINS

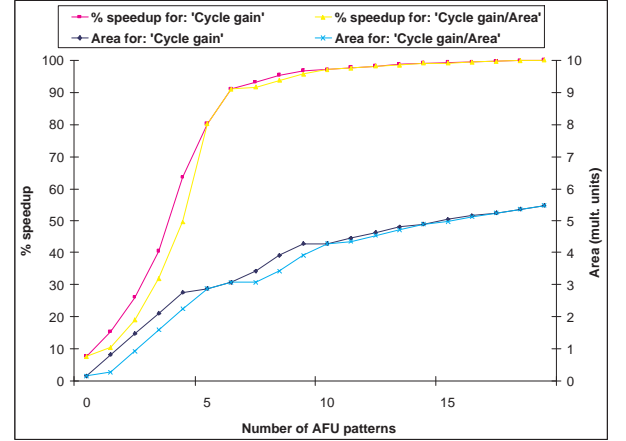


FIGURE 10. Custom instruction selection under priority metrics for *sha* ($N_i/N_o = \infty/\infty$).

TABLE 5. Speedup-AFU area for ‘Cycle gain’/‘Cycle gain/Area’ for the input/output constraint $N_i/N_o = \{8/4\}$.

Benchmark	0.95 × max. speedup	Area (MAU)	At max. speedup	Area (MAU)
<i>adpcmdec</i>	4/4	0.895/0.895	6	0.983
<i>adpcmdec.dlx</i>	11/11	1.123/1.123	17	1.721
<i>adpcmenc</i>	4/4	0.998/0.998	6	1.086
<i>adpcmenc.dlx</i>	16/16	1.475/1.375	22	2.074
<i>crc32.dlx</i>	3/3	0.12/0.12	3	0.12
<i>deraiden</i>	4/4	2.657/2.657	4	2.657
<i>enraidn</i>	3/3	1.949/1.949	3	1.949
<i>fir</i>	4/4	1.398/1.398	5	1.398
<i>fir.dlx</i>	2/2	0.155/0.155	2	0.155
<i>fsme.dlx</i>	9/9	1.143/1.143	11	1.546
<i>fsme.dlx</i>	6/6	1.03/1.03	10	1.65
<i>idea.dlx</i>	40/50	10.23/9.325	69	13.002
<i>mc</i>	5/5	1.824/1.824	7	2.53
<i>mc.dlx</i>	7/7	1.489/1.489	12	2.516
<i>sha.dlx</i>	7/7	1.671/1.671	20	3.378

[33] compilers and the ArchC [34] simulation framework. Functional and cycle-accurate simulators generated by version 1.5.1 of ArchC can be used with YARDstick without any modifications. Most of the YARDstick functionality is also accessible through a cross-platform GUI [7] compatible to recent Tcl/Tk versions (8.5.a5 and newer).

Supported platforms include GNU/Linux (RedHat 9.0), Cygwin and Win32 (Windows/XP SP2) on x86-compatible processors.

6. CONCLUSIONS

YARDstick is a retargetable application analysis and custom instruction generation/selection environment providing a compiler-/simulator-agnostic infrastructure. YARDstick aims in separating design space exploration from compiler/simulator idiosyncrasies. Different compilers/simulators can be plugged-in via file-based interfaces; further, both high- (e.g. ANSI C) and low-level (assembly for an architecture or a virtual machine) input can be analyzed by the infrastructure.

In order to prove the applicability and usefulness of YARDstick in ASIP development, we have evaluated a variety of exploration scenarios on a benchmark set consisting of well-known embedded applications and kernels. In this context, we have investigated effects of the compilation process, such as the selection of the target IR and the impact of register allocation, on the characteristics of the identified hardware extensions. Also, different memory models involving local storage for application-specific functional units were examined and quantified, and for the entire set of applications, custom instructions were generated under different input/output constraints.

REFERENCES

- [1] ARC cores. <http://www.arccores.com>.
- [2] Gonzalez, R. (2000) Xtensa: A configurable and extensible processor. *IEEE Micro*, **20**, 60–70.
- [3] Yiannacouras, P., Steffan, J. G., and Rose, J. (2006) Application-specific customization of soft processor microarchitecture. *Proceedings of the 2006 ACM/SIGDA 14th International Symposium on Field Programmable Gate Arrays*, Monterey, California, USA, February 22–24, pp. 201–210.
- [4] Altera Nios-II home page. <http://www.altera.com/products/ip/processors/nios2/>.
- [5] Xilinx home page. <http://www.xilinx.com>.
- [6] Sirowy, S., Wu, Y., Lonardi, S., and Vahid, F. (2007) Two-level microprocessor accelerator partitioning. *Proc. of the Design, Automation and Test in Europe Conf.*, Nice, France, April 16–20, pp. 313–318.
- [7] Kavvadias, N. and Nikolaidis, S. (2007) YARDstick: Automation tool for custom processor development. *presented at the University Booth of the Design, Automation and Test in Europe Conference (DATE'07)*, April 16–20.
- [8] Clark, N., Zhong, H., Tang, W., and Mahlke, S. (2003) Automatic design of application specific instruction set extensions through dataflow graph exploration. *Int. J. of Parallel Programming*, **31**, 429–449.
- [9] Castro, P., Borin, E., Azevedo, R., and Araujo, G. (2004) Looking for instruction patterns in the design of extensible processors. *Proc. 3rd Wshp. on Application Specific Processors*, Stockholm, Sweden, September.
- [10] Alippi, C., Fornaciari, W., Pozzi, L., and Sami, M. (1999) A DAG based design approach for reconfigurable VLIW processors. *Proc. of the Design, Automation and Test in Europe Conf.*, Munich, Germany, March, pp. 778–779.
- [11] Yu, P. and Mitra, T. (2004) Scalable custom instructions identification for instruction-set extensible processors. *Proc. Int. Conf. on Compilers, Architectures and Synthesis for Embedded Systems*, Washington, DC, USA, September.
- [12] Clark, N., Blome, J. A., Chu, M. L., Mahlke, S. A., Biles, S., and Flautner, K. (2005) An architecture framework for transparent instruction set customization in embedded processors. *Proc. 32nd Int. Symp. on Computer Architecture*, Madison, Wisconsin, USA, June, pp. 272–283.
- [13] Goodwin, D. and Petkov, D. (2003) Automatic generation of application specific processors. *Proc. Int. Conf. on Compilers, Architectures and Synthesis for Embedded Systems*, San Jose, California, USA, October, pp. 137–147.
- [14] Pozzi, L., Atasu, K., and Ienne, P. (2006) Exact and approximate algorithms for the extension of embedded processor instruction sets. *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, **25**, 1209–1229.
- [15] Biswas, P., Dutt, N. L., Pozzi, L., and Ienne, P. (2007) Introduction of architecturally visible storage in instruction set extensions. *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, **26**, 435–446.
- [16] Pothineni, N., Kumar, A., and Paul, K. (2007) Application specific datapath extension with distributed i/o functional units. *Proceedings of the 20th International Conference on VLSI Design (VLSI Design 2007), Sixth International Conference on Embedded Systems (ICES 2007)*, Bangalore, India, January 6–10, pp. 551–558.
- [17] Pattlib. <http://www.lsc.ic.unicamp.br/pattlib/>.
- [18] GCC. <http://gcc.gnu.org>.
- [19] Atasu, K., Pozzi, L., and Ienne, P. (2003) Automatic application-specific instruction-set extensions under microarchitectural constraints. *Proc. 40th ACM/IEEE Design Automation Conference*, June, pp. 256–261.
- [20] Kavvadias, N. and Nikolaidis, S. (2005) Automated instruction-set extension of embedded processors with application to MPEG-4 video encoding. *Proc. 16th Int. Conf. on Application-specific Systems, Architectures and Processors*, July 23–25, pp. 140–145.
- [21] The VFLib graph matching library, version 2.0. <http://amalfi.dis.unina.it/graph>.
- [22] VCG. <http://rw4.cs.uni-sb.de/sander/html/gsvcg1.html>.
- [23] GraphViz. <http://www.graphviz.org>.
- [24] CDFG toolset. <http://poppy.snu.ac.kr/CDFG/cdfg.html>.
- [25] AGG. <http://tfs.cs.tu-berlin.de/agg/>.
- [26] SALTO. <http://www.irisa.fr/caps/projects/Salto/>.
- [27] Machine-SUIF research compiler. <http://www.eecs.harvard.edu/hube/software/>.
- [28] Raiden: An alternative to TEA block cipher. <http://raiden-cipher.sourceforge.net>.
- [29] Bonzini, P. and Pozzi, L. (2006) Code transformations strategies for extensible embedded processors. *Proc. of the 2006 Int. Conf. on Compilers, Architectures and Synthesis for Embedded Systems*, Seoul, Korea, October 22–25, pp. 242–252.
- [30] ACE – CoSy compiler development system. <http://www.ace.nl>.
- [31] Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, USA.
- [32] Leupers, R., Karuri, K., Kraemer, S., and Pandey, M. (2006) A design flow for configurable embedded processors based on optimized instruction set extension synthesis. *Proc. of the Design, Automation and Test in Europe Conf.*, Messe Munich, Germany, March 6–10.
- [33] The COINS project. <http://www.coins-project.org>.
- [34] The ArchC resource center. <http://www.archc.org>.
- [35] Yu, P. and Mitra, T. (2004) Characterizing embedded applications for instruction-set extensible processors.

- Proc. 41th ACM/IEEE Design Automation Conference*, San Diego, CA, USA, June, pp. 723–728.
- [36] Guthaus, M. R., Ringenberg, J. S., Ernst, D., Austin, T. M., Mudge, T., and Brown, R. B. (2001) MiBench: A free, commercially representative embedded benchmark suite. *Proc. of the 4th annual IEEE Int. Wshp. on Workload Characterization*, Austin, Texas, USA, December.
- [37] Cong, J., Fan, Y., Han, G., and Zhang, Z. (2004) Application-specific instruction generation for configurable processor architectures. *Proc. 12th Int. Symp. on Field Programmable Gate Arrays*, Monterey, California, USA, February, pp. 183–189.
- [38] Sander, G. (1994) Graph layout through the VCG tool. *Proc. DIMACS Int. Workshop on Graph Drawing*, Berlin, Germany, pp. 194–205.
- [39] Gansner, E. R., Koutsofios, E., North, S. C., and Vo, K.-P. (1993) A technique for drawing directed graphs. *IEEE Trans. Softw. Eng.*, **19**, 214–230.
- [40] Briggs, P. and Harvey, T. (1994) Multiplication by integer constants. Technical report. Rice University.
- [41] Lee, L. H., Moyer, W., and Arends, J. (1999) Instruction fetch energy reduction using loop caches for embedded applications with small tight loops. *Proc. Int. Symp. on Low Power Electronics and Design*, San Diego, CA, August.
- [42] Hennessy, J. and Patterson, D. (1994) *Computer Organization and Design: The Hardware/Software Interface*. Morgan Kaufmann Publishers.